J. R. Statist. Soc. A (1999)
162, *Part* 3, *pp.* 363–382

Accounting for inaccuracies in population counts and case registration in cancer mapping studies

Nicky Best and Jon Wakefield

Imperial College School of Medicine, London, UK

[Received December 1998. Revised June 1999]

Summary. Disease mapping studies summarize spatial and spatiotemporal variations in disease risk. This information may be used for simple descriptive purposes, to assess whether health targets are being met or whether new policies are successful, to provide the context for further studies (by providing information on the form and size of the spatial variability in risk) or, by comparing the estimated risk map with an exposure map, to obtain clues to aetiology. There are well-known problems with mapping raw risks and relative risks for rare diseases and/or small areas since sampling variability tends to dominate the subsequent maps. To alleviate these difficulties a multilevel modelling approach may be followed in which estimates based on small numbers are 'shrunk' towards a common value. In this paper we extend these models to investigate the effects of inaccuracies in the health and population data. In terms of the health data we consider the effects of errors that occur due to the imperfect collection procedures that are used by disease registers. For cancers in particular, this is a major problem, with case underascertainment (i.e. undercount) being the common type of error. The populations that are used for estimating disease risks have traditionally been treated as known quantities. In practice, however, these counts are often based on sources of data such as the census which are subject to error (in particular underenumeration) and are only available for census years. Intercensual population counts must consider not only the usual demographic changes (e.g. births and deaths) but migration also. We propose several approaches for modelling population counts and investigate the sensitivity of inference to the sizes of these errors. We illustrate the methods proposed using data for breast cancer in the Thames region of the UK, and we compare our results with those obtained from more conventional approaches.

Keywords: Case underascertainment; Census underenumeration; Disease mapping; Multilevel models

1. Introduction

The aims of spatial epidemiological studies include the simple description of geographical variability in risk, the investigation of the sources of the spatial and non-spatial components of this variability and cluster investigations, either with or without any specific putative source. To carry out such studies we require data on health events and on an appropriate 'denominator'. The methodology that is used depends on whether exact locations are available (point data) or whether aggregation has been carried out (count data). Count data are more likely to be routinely available and it is these that we consider here. In general, each of the health, population and exposure data may be subject to error. Wakefield and Elliott (1999) have provided a detailed discussion of the sources of these errors. Errors-in-variables modelling of exposure data has been considered by various researchers including Bernardinelli *et al.* (1997) and Jordan *et al.* (1997), but inaccuracies in the health and population data have received less

Address for correspondence: Jon Wakefield, Small Area Health Statistics Unit, Department of Epidemiology and Public Health, Imperial College School of Medicine, St Mary's Campus, Norfolk Place, London, W2 1PG, UK.
E-mail: j.c.wakefield@ic.ac.uk

© 1999 Royal Statistical Society 0964–1998/99/162363

attention in a disease mapping context. In this paper we consider the modelling and effects of specific errors that may occur in the health and population data. So far as the health data are concerned we focus on cancer incidence data and the modelling of inaccuracies that occur in case registration, e.g. double counting and under-registration. Population information is often taken from the census (which in the UK is carried out every 10 years). Such data sources only provide *estimates* of the population, however; in particular underenumeration (undercount) is a major problem (Simpson *et al.*, 1995). Here we do not consider underenumeration but, rather, the problem of dealing with the intercensual years in which populations may change due to births and deaths and also to migration. The inaccuracies in health and population data are not likely to be spatially neutral (i.e. the probability of an error is not geographically constant) and so observed differences in area-specific relative risk estimates may reflect anomalies in the data, as opposed to real differences. Here we provide models for each of these problems in the context of a disease mapping study.

The structure of this paper is as follows. In Section 2 we describe the data that we shall use to illustrate our methods. These data concern the incidence of breast cancer in the Thames region in the years 1981–1991; in particular we describe the manner of data collection and sources of error within the health and population data. In Section 3 we present a statistical model that may be used to analyse areal count data and in Section 4 we extend this model to account for inaccuracies in the data and carry out several analyses using a range of models. Section 5 considers the effect of introducing spatial effects, and Section 6 contains a concluding discussion.

2. Breast cancer data

We illustrate our methods via a disease mapping study that was carried out to investigate sources of variability in the relative risk of breast cancer in women aged 15 years and over, at electoral ward level, in the Thames region of the UK. There are 2136 wards in the study region with a total of 74875 registered cases over the period 1981–1991.

2.1. Registrations

To motivate our models we briefly review the complex cancer registration system that operates in England and Wales. Swerdlow (1986) has provided a more detailed description; see also Gulliford *et al.* (1993). England and Wales are covered by a number of regional cancer registries who voluntarily supply the Office for National Statistics (ONS) with cancer registration data. (During the study period, the ONS were known as the Office of Population Censuses and Surveys.) Data available for each registration include the cancer site and the date of birth, sex and postcode of residence (at the time of diagnosis) of the individual. These data are collected by the registries in a variety of ways. For example some employ peripatetic clerks, others use hospital record staff to extract data and others rely on the computer systems of other organizations such as hospitals and pathology laboratories. Only rarely have studies been carried out to estimate the probability that a new case is registered, i.e. the completeness (see for example Hawkins and Swerdlow (1992)). Other issues that must be considered include the accuracy of diagnoses and encoding (including consistency of clerical staff and the effect of training programmes), late registrations, deletions and amendments, constancy of clinical and pathological coding and changes in coding systems. Each registry consists of a number of district health authorities (DHAs) and the aforementioned issues may result in varying completeness between DHAs, as well as between registries. The DHA effect is more difficult to model since, although each ward may be mapped to a DHA, wards on the boundary of DHA

catchment areas may contain populations who attend more than one hospital. Differences in ascertainment must be addressed in any geographical study since when maps are produced apparent highs and lows may simply correspond to registration anomalies.

Once the data have been received, the ONS carry out validation and checks for duplicates before summaries are released. Included in this process is the cross-checking of registrations with the National Health Service Central Register. In recent years 96% completeness was found (Office for National Statistics, 1997). All deaths within England and Wales are recorded with the ONS and deaths are linked with the registration data.

The proportion of individuals who were recorded by a regional registry as dying of a particular cancer for whom no other information could be found, the so-called 'death certificate only' (DCO) registrations, provides one indicator of incomplete case ascertainment.

Before 1985, the Thames region comprised three separate registries, namely the North West, North East and South registries; in 1985 the three registries amalgamated. This amalgamation led to an overall improvement in the registration process in the Thames region.

One measure of the level of ascertainment is provided by calculating the ratio of mortality to incidence for a particular cancer site. If all cases are registered (and the deaths are recorded) then this ratio will approximately equal the probability of death given cancer at a particular site. For breast cancer this probability is approximately 0.5 (Parkin *et al.*, 1994). Fig. 1 shows this ratio for each of the three regions of Thames for the study period. We see that for the North East and North West regions the ratios were in excess of 0.5 before 1985. This could be due to a change in risk factors or a change in registration procedures. Given the amalgamation in 1985, the latter possibility seems the more likely explanation.

Swerdlow and dos Santos Silva (1993), in their cancer atlas of England and Wales, used age-adjusted odds ratios for all cancers except childhood leukaemia to account for case under-ascertainment. A case–control method was used in which a weighted sample of other cancers was used as the control group. This method depends on the relative underascertainment within a county (the areal units at which maps were displayed) compared with the study region being constant across tumour sites. The weighting procedure ensured that the maximum contribution of any one cancer site was taken to be less than 7% to avoid controls being dominated by a few common cancers. We describe a similar method in Section 4. It is more usual for cancer atlases to be produced for mortality for which, in general, registration problems are less severe. There is still an appreciation of potential inaccuracies, however; see for example Pickle *et al.* (1996), pages 3–5.

Breast cancer screening was introduced in 1989 and hence we would expect to see a subsequent apparent increase in incidence in lower age groups (followed in later years by a decline in older age groups, assuming approximately constant risk across years). In Fig. 1 we see that the mortality/incidence ratio drops in 1989 owing to increased incidence.

York *et al.* (1995) and Bray and Wright (1998) considered the modelling of underascertainment for congenital malformations. The methods described in these papers were developed for situations in which there were two sources of registration information, however, and so are not directly applicable here.

2.2. Population counts

So far as population data are concerned the raw census counts are available by ward and by 5-year age bands for 1981 and 1991. For 1991 these census counts have been adjusted by the 'Estimating with confidence' project (Simpson *et al.*, 1995) for underenumeration, to move students from their term time address to their home address and to provide a mid-year estimate (as opposed to a census date estimate). The resulting counts remain an estimate of

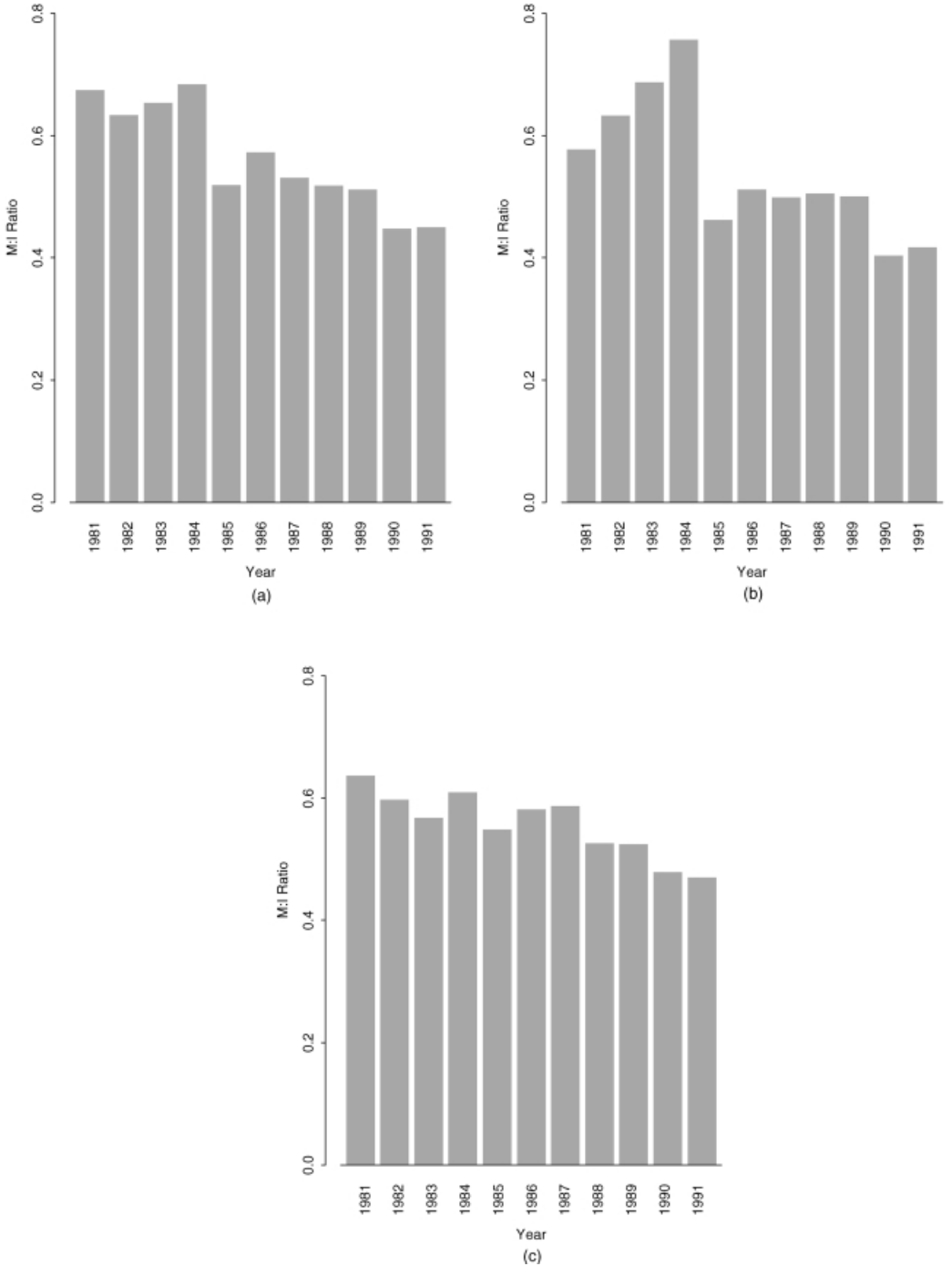

Fig. 1. Ratio of counts of incidence to counts of mortality for breast cancer for three Thames regions over the period 1981–1991 (before 1985 there were three separate registries in North West, North East and South Thames; from 1985 they were amalgamated — if the incidence and mortality data are complete the ratio should be approximately 0.5): (a) North East Thames region; (b) North West Thames region; (c) South Thames region

the 'true' population but are the best set of counts that are available nationally. In 1981 the underenumeration at census was estimated to be 0.5% (Office for National Statistics, 1997). The third set of data that we utilize are the Registrar General's mid-year estimates for the years 1982–1990 which are available by 5-year age bands at the level of local authority district (LAD). There are 96 of these LADs in the Thames region with 8–45 wards per LAD (median 22). The mid-year estimates roll the 1981 census figures forwards using information on births, deaths and migration (Office of Population Censuses and Surveys, 1991).

In our disease mapping context, when population data as described above are utilized, we must consider the following.

(a) The census is not entirely accurate and in particular suffers from underenumeration which will not be spatially neutral.
(b) For intercensual years the Registrar General's LAD estimates are not exact, and if smaller areas than these districts are considered (e.g. wards) then these counts must be apportioned, introducing further error.

In a health context, migration not only affects the raw counts but also the exposure to area-defined risk factors that the migrant population experiences. For example, individuals may be exposed to some source of pollution but then move to another area before the disease develops, hence diluting the observed effect in the area containing the source of pollution.

We note that the effects of migration will be greatest for small areas; within larger areas individuals tend to migrate *within* the same area (Richardson, 1992). So, when disease maps are examined, interesting features may simply reflect population denominator problems.

3. Basic model

In this section we specify the basic disease mapping model that we utilize. Clayton and Bernardinelli (1992) and Mollié (1996) have provided reviews of the aims and methodology of disease mapping. We first establish notation: let N_{iat} and Y_{iat} denote respectively the population at risk and the number of registered breast cancer cases in area i, age band a and calendar year t. For the breast cancer data, i therefore indexes the 2136 wards and a the 14 5-year age bands 15–19, . . ., 80–84 years, plus one age band for those over 84 years old, with t ranging over 1981, . . ., 1991. For rare and non-infectious diseases we may then assume that

$$Y_{iat} \underset{\text{IID}}{\sim} \text{Poisson}(N_{iat} p_{iat}), \tag{1}$$

where p_{iat} is the probability of disease in area i, age group a and calendar year t. We note that this distributional assumption will often be inaccurate owing to the effect of unmeasured risk factors which, in particular, will lead to extra-Poisson variability. One method of acknowledging this variability is via the introduction of random effects, as described below. Model (1) also assumes that the risk for a given age by year stratum is constant *within* each area; this is clearly an approximation to the 'true' underlying relative risk surface. In particular, the analysis is open to the possibility of the *ecological fallacy* in which incorrect inference is made because an area level relationship is only relevant to the individuals within that area under very strict conditions (Richardson, 1992). The disease mapping approaches that are currently available are all vulnerable to this problem, and the possibility of ecological bias should always be borne in mind at the interpretation stage.

Conventionally it is then assumed that

$$p_{iat} = \theta_{it} p_{at}, \tag{2}$$

where p_{at} are stratum-specific reference rates. These are often estimated marginally from the whole study region via

$$\hat{p}_{at} = \sum_i Y_{iat} \Big/ \sum_i N_{iat} = Y_{+at}/N_{+at} \tag{3}$$

and are treated as known. Model (2) implies the very strong assumption that the effect of being in area i in year t is to multiply all the stratum-specific risks by a common *relative risk* θ_{it}. For the breast cancer data we assessed this proportionality assumption informally via the examination of plots of maximum likelihood estimates of $\hat{p}_{iat}$ *versus* $\hat{p}_{at}$ for areas that had sufficient cases to produce reliable estimates. The relationships were approximately linear in most areas, suggesting that the proportionality assumption is reasonable for these data.

The Poisson assumption (1) combined with the proportionality assumption (2) then allows summation over the stratum to give

$$Y_{it} \sim \text{Poisson}(E_{it}\theta_{it}), \tag{4}$$

where $Y_{it} = \Sigma_a Y_{iat}$ and

$$E_{it} = \sum_a N_{iat}\hat{p}_{at}. \tag{5}$$

Inaccuracies in the population data will be reflected in the expected numbers E_{it} both through the counts for the small areas and the stratum-specific probabilities. The implementation of model (3)–(5) achieves a significant reduction in the number of data items compared with model (1). If this is not an issue then the stratum-specific probabilities may be estimated simultaneously with the small area relative risk estimates (see Clayton (1996)). The maximum likelihood estimator of the relative risk θ_{it} is the standardized morbidity ratio (SMR) Y_{it}/E_{it}.

In disease mapping applications, following Clayton and Kaldor (1987), it has become the norm to specify a hierarchical model for the data. The expected *similarity* of relative risks is modelled via the incorporation of random effects. Specifically the relative risks may be modelled via

$$\log(\theta_{it}) = \mu_t + X_i^{\mathrm{T}}\beta + \upsilon_i \tag{6}$$

where X_i denotes a $k \times 1$ vector of ward level explanatory variables and β a $k \times 1$ vector of regression coefficients, and the $\upsilon_i \sim_{\text{IID}} N(0, \sigma_\upsilon^2)$ model global similarity. The introduction of the random effects υ_i yields a marginal distribution for Y_i that is overdispersed relative to the Poisson distribution; the parameters υ_i may be interpreted as due to unmeasured risk factors that are common to the individuals of area i (but do not have spatial structure). For the observed explanatory variables, we take X_i to be the Carstairs deprivation index (Carstairs and Morris, 1991) of ward i evaluated in 1991. This index is calculated from census variables concerning overcrowding, access to a car, the social class of the head of household and unemployment. These combine to give a univariate continuous measure, with high values indicating greater deprivation. Using this index we may take some account of socioeconomic status, which is well known to be a strong predictor of disease (e.g. Jolley *et al.* (1992)). Since $k = 1$ in all our models, from this point onwards we drop the transpose on X_i.

Besag *et al.* (1991) suggested that model (6) may be augmented by the addition of random effects u_i that have spatial structure since, in general, we would expect at least some of the residual risk to exhibit spatial dependence. For now, we do not consider this extension, since our aim is to illustrate the effects of inaccuracies in the data; however, we return to this point in Section 5.

A great benefit of model (6) is that it results in relative risk estimates that have 'borrowed strength' from the totality of areas and hence are not subject to the great instability

that raw SMRs based on small expected numbers may exhibit (Clayton and Kaldor, 1987).

A Bayesian approach completes the model by specifying prior distributions for μ_t, β and σ_v^2. Throughout we have used normal priors $N(0, 1000)$ for fixed effects and gamma priors Ga(0.5, 0.0005) for random-effects precisions (σ_v^{-2}). For a justification of the latter see Kelsall and Wakefield (1999). All models were fitted using Markov chain Monte Carlo simulation algorithms implemented in the WinBUGS software (Spiegelhalter *et al.*, 1998). Posterior summaries were based on samples of 10000 after discarding a 1000-iteration burn-in, convergence having been checked via an informal assessment of trace and quantile plots.

Depending on the aim of the study, the inference may focus on a variety of parameters. Ecological correlation studies will focus on the regression parameters β. To place spatial epidemiological studies in context we may examine the size of σ which corresponds (approximately) to the standard deviation of the residual relative risks $\exp(v_i)$. We may be interested in the area level relative risks θ_{it} for allocating health services and investigations of unexplained variability will concentrate on the residual relative risks $\exp(v_i)$.

We first fitted model (3)–(6) separately to the data in the periods 1981–1984 and 1985–1991, where the E_{it} have been calculated using a set of age- and calendar-year-specific probabilities (3) calculated for the whole region using 1981 population census counts for years before 1986 and 1991 census counts for 1986 onwards. Fig. 2 shows a map of the ratio of the posterior means of the residual relative risk estimates $\exp(v_i)$ in the 1981–1984 and 1985–1991 periods for breast cancer by ward in the Thames region. The predominance of ratios less than 1 in the northern part of the region is likely to be due to the case ascertainment that was seen in Fig. 1 in North Thames before the amalgamation with the South Thames cancer registry in 1985 (see Section 2.1). Of course without additional information it is impossible to determine the extent to which this figure reflects differences in risk factors that are common to the registries and case underascertainment. The inaccuracies in the population counts discussed in Section 2.2 may also be having an influence.

Fig. 3(a) displays the residual relative risks $\exp(v_i)$ for each ward in the 1981–1984 period plotted against the corresponding estimates in the 1985–1991 period; these were the quantities whose ratios were displayed in Fig. 2. Although we would expect to see differences in these two time periods due to changes in the case mix, the lack of consistency between these estimates again suggests that problems in the data may be important here.

4. Extended model

4.1. Registrations

In this section we first describe how we propose to account for case ascertainment problems. *A priori* we would expect that the principal differences in ascertainment will occur between registries though there will be within-registry differences, at the DHA level, as well. Consequently we introduce some new notation: let $r[i]$ and $d[i]$ denote respectively the registry and DHA within which ward i lies.

To investigate the registry effects as a function of year we fitted the model given by equations (3)–(5), and replacing equation (6) by

$$\log(\theta_{it}) = \mu + X_i\beta + \eta_{r[i]t} + v_i \tag{7}$$

where $v_i \sim_{\text{IID}} N(0, \sigma_v^2)$ and the fixed effects $\eta_{r[i]t}$ estimate registry effects by year. Fig. 4 displays the posterior means of $\eta_{r[i]t}$ for the $r[i] = 1, 2, 3$ regions by year. We see that there is greater

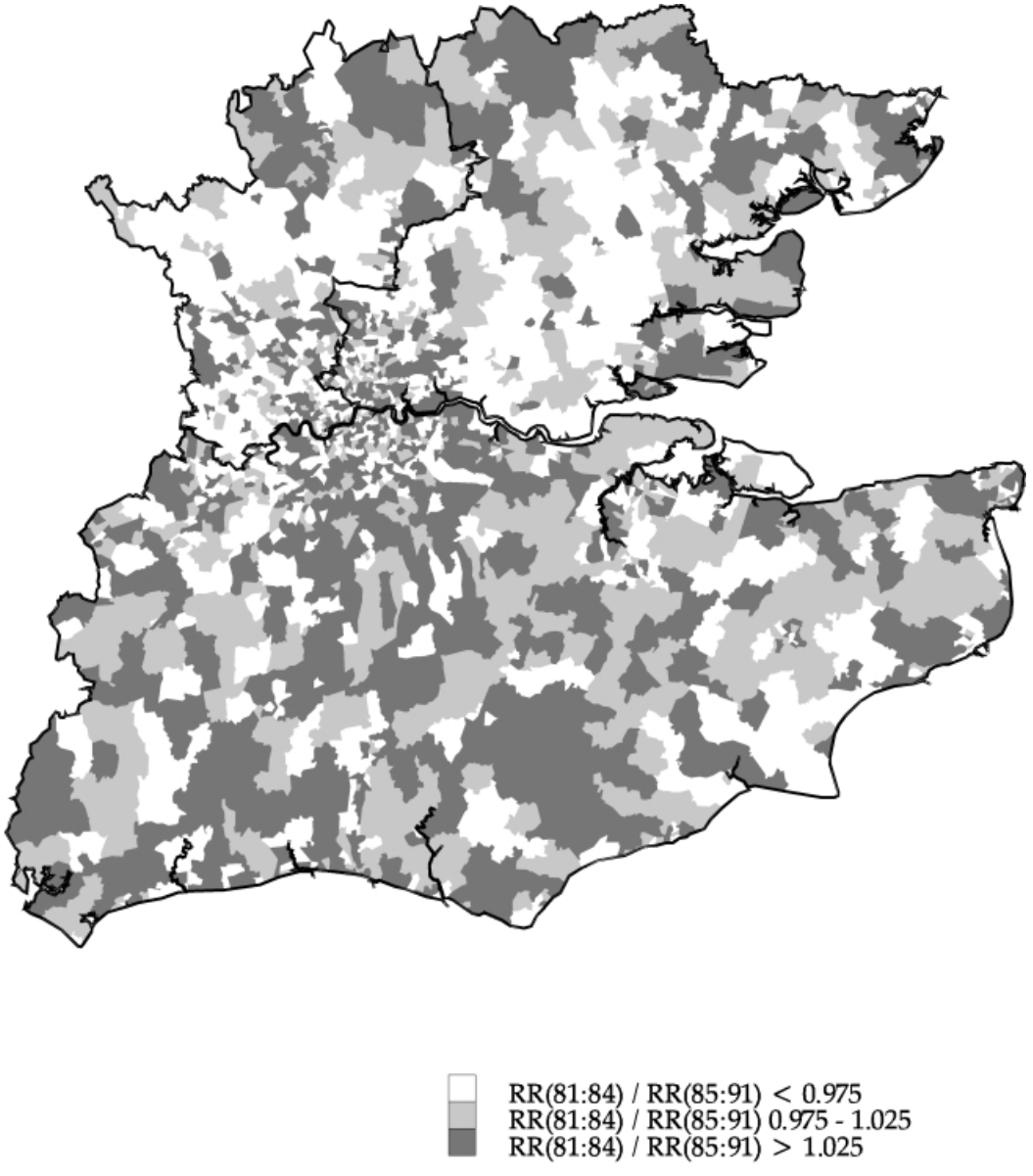

Fig. 2. Ratio of the ward level residual relative risks $\exp(v_i)$ in 1981–1984 and 1985–1991: ———, registry boundaries before the amalgamation in 1985

variability before the amalgamation of the registries in 1985. In particular before 1985 the relative risk in the southern registry is greater. Again, this is probably a reflection of the greater ascertainment that apparently occurred in this region (see Fig. 1), rather than a true elevated risk of breast cancer.

We may similarly investigate differences between DHAs by fitting the model

$$\log(\theta_{it}) = \mu + X_i\beta + \eta_{r[i]t} + \phi_{d[i]t} + v_i, \tag{8}$$

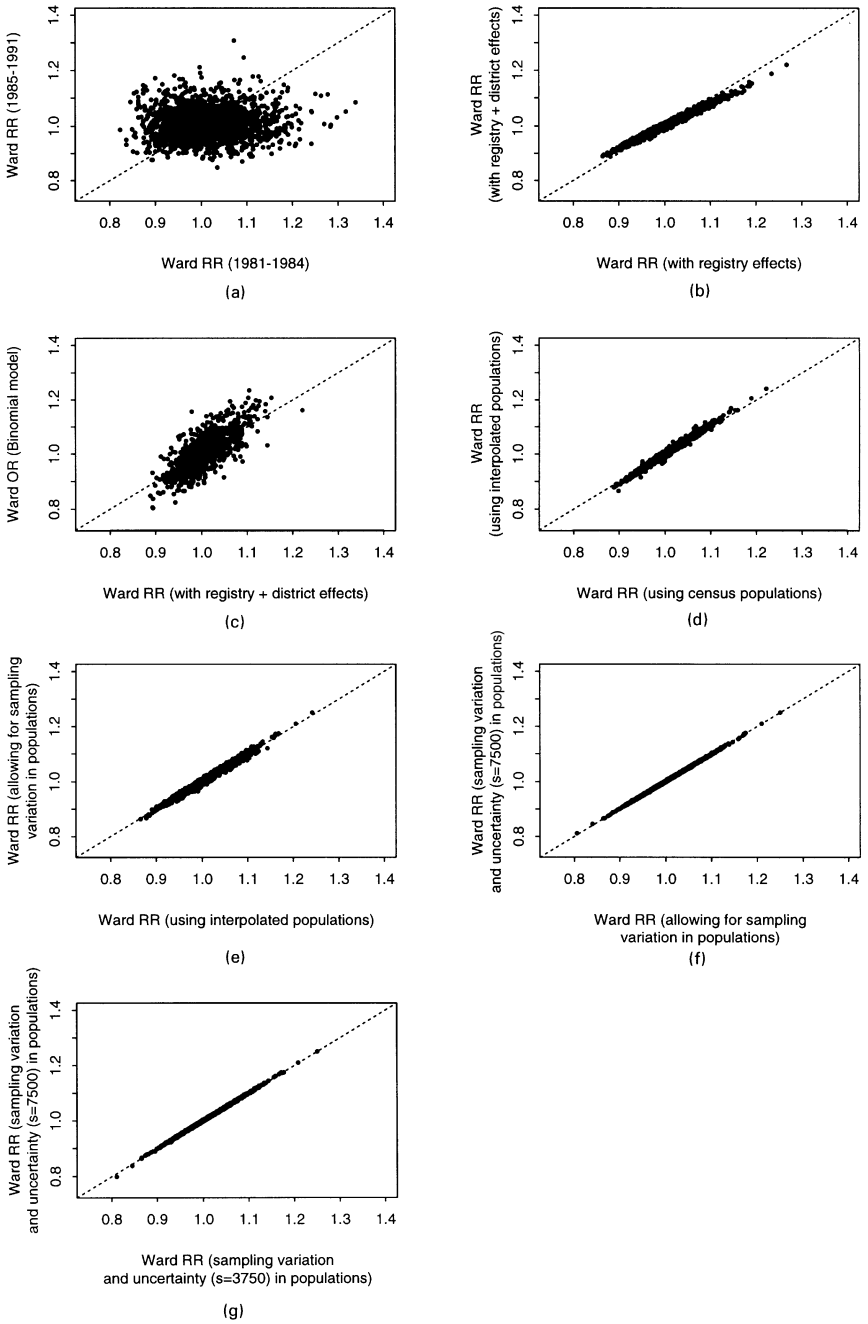

Fig. 3 (a) Residual relative risk (RRR) estimates by ward in 1981–1984 *versus* estimates in 1985–1991; (b) RRR estimates with adjustment for registry *versus* estimates with adjustment for registry and district; (c) RRR estimates with adjustment for registry and district *versus* estimates from the proportional incidence binomial model; (d) RRR estimates using census populations *versus* estimates using interpolated populations; (e) RRR estimates using interpolated populations *versus* estimates using multinomial ward populations; (f) RRR estimates using multinomial ward populations *versus* estimates using Dirichlet–multinomial ward populations; (g) RRR estimates using Dirichlet–multinomial ward populations with prior variance proportional to 7500^{-1} *versus* estimates using Dirichlet–multinomial ward populations with prior variance proportional to 3750^{-1}

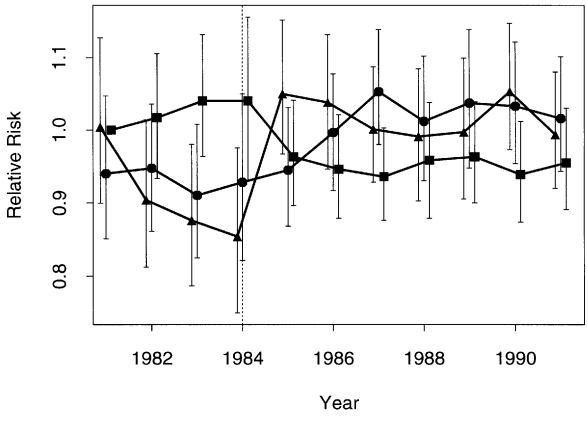

Fig. 4. Registry effects (with 95% interval estimates) for the study period (the registries were amalgamated at the beginning of 1985): ▲, North East Thames region; ●, North West Thames region; ■, South Thames region

with DHA effects $\phi_{d[i]t} \sim_{\text{IID}} N(0, \sigma_t^2)$. We choose to model the DHA effects as random effects because some of these contain sparse data and so benefit from being smoothed to the overall mean. Fig. 3(b) shows the residual ward level relative risk estimates $\exp(v_i)$ from the model with registry effects (7) plotted against the estimates from the model with registry *and* district effects (8). The agreement between these two sets indicates that the dominant contribution to the relative risk is made, as we would expect, by the registry effects. The range of the residual relative risks is narrower after district effects have been added since we have accounted for slightly more of the variability. The posterior medians of σ_t for the preamalgamation years $t = 1981, \ldots, 1984$ are given by

$$0.09, 0.09, 0.08, 0.12,$$

whereas for the post-amalgamation years $t = 1985, \ldots, 1991$ they are given by

$$0.05, 0.05, 0.03, 0.06, 0.08, 0.04, 0.04,$$

and so we see that the variability narrows following amalgamation, suggesting that cancer registration procedures became more homogeneous after 1984.

We now describe a second approach that follows more closely the method of Swerdlow and dos Santos Silva (1983); see also Breslow and Day (1987). Following the notation for the breast cancer counts, let Z_{iat} denote the number of cases of all other cancers (excluding cancer of the uterus and ovary for reasons given later) in area i, age band a and year t. We consider the model

$$Y_{iat} \underset{\text{IID}}{\sim} \text{Poisson}\{N_{iat} \exp(\gamma_{at} + X_i\beta + \eta_{r[i]t} + \phi_{d[i]t} + v_i)\}$$

and

$$Z_{iat} \underset{\text{IID}}{\sim} \text{Poisson}\{N_{iat} \exp(\gamma'_{at} + X_i\beta' + \eta_{r[i]t} + \phi_{d[i]t} + v'_i)\}.$$

Hence we have assumed that the stratum- (age–year) and area-specific risks differ for breast cancer *versus* all other cancers combined (via the γ and γ' fixed effects, and the v and v' random effects respectively) but the registry and DHA effects are the same across all tumour sites combined and for breast cancer. This assumption deserves some discussion; the

proportion of underascertainment will be related to the proportion of DCO cases but the relationship is not likely to be constant across tumour sites since the mortality-to-incidence ratio is not constant across sites. For example, for lung cancer, survival is very poor and so if all cases were registered as DCO cases the underascertainment would be small. This contrasts with breast cancer for which long-term survival is approximately 50% (see Fig. 1) and so if all cases were registered as DCO cases the underascertainment would be greater than for lung cancer since many cases would die from other causes which may lead to the registry's not being notified and the case would be lost. Hence the assumption depends critically on whether the proportions of DCO cases for all tumour sites other than breast (and uterus and ovarian) are consistent across the registries and DHAs, relative to the proportion for breast cancer. Obviously it would be preferable to obtain information on these proportions from the registries, but unfortunately such information is not available to us.

We now condition on the total cancers as the denominator to give

$$Y_{iat} | Y_{iat} + Z_{iat} \underset{\text{IID}}{\sim} \text{binomial}(Y_{iat} + Z_{iat}, R_{iat}) \tag{9}$$

where

$$\text{logit}(R_{iat}) = \gamma^*_{at} + X_i\beta^* + v^*_i,$$

and $\gamma^*_{at} = \gamma_{at} - \gamma'_{at}$ is the difference in stratum risks, $v^*_i = v_i - v'_i$ is the difference in area effects for breast cancer *versus* other cancers and $\beta^* = \beta - \beta'$ measures the difference between the effects of the Carstairs deprivation measure on breast cancer *versus* all other cancers. This approach therefore eliminates the denominator and registry or DHA problems, the latter under the assumption that the underascertainment across all other cancer sites combined is the same as for breast cancer. The interpretation of the v^*_i is of the difference in unobserved risk factors acting on breast *versus* all other cancers within a registry. This interpretation indicates why we excluded cancer of the uterus and ovary from the control group since these are thought to have a similar aetiology to breast cancer. A computational disadvantage of this approach is that the incidence of breast cancer relative to all cancers is no longer small and so we cannot approximate model (9) by a Poisson distribution which would allow us to combine cases across the stratum. A further difficulty is that this approach cannot, in general, be followed if it is an ecological study that is being carried out. This is because it is $\beta - \beta'$ that is being estimated and not β; the latter can only be estimated if knowledge of β' is available.

Fig. 3(c) displays the residual ward level relative risk estimates obtained from the Poisson model with registry and district effects *versus* the corresponding odds ratio estimates from the proportional binomial model just described. The odds ratios $\exp(v^*_i)$ correspond to

$$\frac{\text{odds of breast cancer } \textit{versus} \text{ all other cancers in ward } i}{\text{average odds of breast cancer } \textit{versus} \text{ all other cancers across the Thames region}}.$$

The Poisson and binomial models are modelling different quantities (the residual relative risk and residual odds ratio respectively) and so the estimates are not directly comparable (although the rank ordering should be unaffected). The strength of the agreement is encouraging as it suggests that both models are, at least to some extent, accounting for underascertainment.

4.2. Population counts

In this section we shall always include registry and DHA effects, i.e. model (8) is our starting point. Recall that the N_{iat} are the 'true' population counts at ward level. In what follows we

shall treat the census estimates as the true counts and consider three models of increasing complexity for the intercensual years. For notational convenience we consider a single LAD, at which the Registrar General's counts are available, without introducing a further subscript. We let N_{+at}, $t = 1982, \ldots, 1990$, denote the Registrar General's set of mid-year counts for this LAD which is assumed to contain I wards. Let π_{iat} denote the true proportion of the stratum a population contained in the LAD that is in ward i and year t.

4.2.1. *Model A*

The first model that we consider assumes that the apportionment probabilities are known. In the years of the census (years $t = 1981, 1991$) we obtain $\pi_{iat} = N_{iat}/N_{+at}$ directly from the census. For intercensual years we propose the simple logistic interpolation

$$\text{logit}(\pi_{iat}) = b_{ia} + c_{ia}t, \tag{10}$$

for $t = 1982, \ldots, 1990$. We then take $N_{iat} = N_{at}\pi_{iat}$.

4.2.2. *Model B*

We now acknowledge the sampling variability of the counts within an LAD via the multinomial model:

$$N_{1at}, \ldots, N_{Iat}|\pi_{at} \sim \text{multinomial}_I(N_{+at}, \pi_{at}) \tag{11}$$

where $\pi_{at} = (\pi_{1at}, \ldots, \pi_{Iat})^{\mathrm{T}}$.

4.2.3. *Model C*

We finally account for the uncertainty in the apportionment probabilities by assuming a Dirichlet prior distribution for π_{at}:

$$(\pi_{1at}, \ldots, \pi_{Iat})|q_{at}, s_{at} \sim \text{Dirichlet}_I(s_{at}q_{1at}, \ldots, s_{at}q_{Iat}) \tag{12}$$

where $q_{at} = (q_{1at}, \ldots, q_{Iat})^{\mathrm{T}}$ is the mean of this distribution, which we take to be the interpolated probabilities, i.e.

$$E(\pi_{iat}|q_{at}, s_{at}) = q_{iat} = \frac{\exp(b_{ia} + c_{ia}t)}{1 + \exp(b_i + c_{ia}t)}. \tag{13}$$

The variance is given by

$$\text{var}(\pi_{iat}|q_{at}, s_{at}) = \frac{q_{iat}(1 - q_{iat})}{s_{at} + 1}$$

and so s_{at} controls the (inverse) variance of the prior.

This model leads to a Dirichlet–multinomial marginal distribution for the true counts which has mean

$$E(N_{iat}|q_{at}, s_{at}) = N_{+at}q_{iat}$$

and variance

$$\text{var}(N_{iat}|q_{at}, s_{at}) = N_{+at}q_{iat}(1 - q_{iat})\frac{N_{+at} + s_{at}}{1 + s_{at}}.$$

So, for each year and for each LAD and age group, the ward populations follow an over-dispersed multinomial distribution, with the overdispersion occurring because we do not know the underlying ward apportionment probabilities but we assume that they follow a Dirichlet distribution.

We now describe how the parameters s_{at} were chosen. From the 1991 census there is information at ward level on migration within the previous year which we can use to obtain an estimate of each of the ward level populations in 1990. These estimates provide a new set of apportionment probabilities for each ward in 1990, denoted q'_{i1990}. We had to take $q'_{ia1990} = q'_{i1990}$ for all a since the census migration data are not available by age strata. We therefore also recalculated interpolated proportions at ward level, i.e. q_{i1990}, analogous to those in equation (13), and assumed $s_{at} = s_t$ for all a. Following model (12)–(13), we assumed a Dirichlet prior for $\pi_{1990} = (\pi_{1,1990}, \ldots, \pi_{I,1990})^{\mathrm{T}}$ with expectation taken as the interpolated values q_{i1990}. Marginally, each apportionment probability then has a beta distribution

$$\pi_{i1990}|q_{1990}, s_{1990} \sim \text{beta}\{s_{1990}q_{i1990}, s_{1990}(1 - q_{i1990})\}. \tag{14}$$

For various choices of s_{1990} we constructed 95% intervals from the marginal distributions (14) for each ward. We then chose s_{1990} so that approximately 95% of these intervals for π_{i1990} contain q'_{i1990}. Following this procedure gave $s_{1990} = 7500$.

To reflect the fact that errors will increase the further that we move from the census we would expect s_t to be a decreasing function of distance from the nearest census. Unfortunately we do not have data to inform a possible form for this relationship; hence we take $s_t = s_{1990}$ for $t = 1982, \ldots, 1989$. We acknowledge that this process is not rigorous but it is consistent with our aim of addressing the sensitivity of inference to inaccuracies in the counts.

Fig. 5 shows a graphical representation of the final Poisson model including district and registry effects to adjust for case under-registration and uncertain apportionment probabilities to adjust for errors in the population counts. This directed acyclic graph (DAG) illustrates the relationship between the elements of the model, in particular the conditional independences. Oval nodes represent random variables (whether observed or not), rectangular nodes fixed quantities, and logical dependences are shown as broken arrows; see Spiegelhalter (1998) for an introduction to, and application of, DAGs.

Since the expected counts are a function of the N_{iat}, we would ideally allow these to be random in the full probability model (as represented in Fig. 5) and thus allow feed-back from the Poisson portion of the model. Unfortunately this model is very computationally expensive to implement and so we instead make the following simplifications. We carry out a separate analysis of the counts data and obtain a sample from the posterior distribution of the N_{iat} which we then use within the analysis of the health data. This approximation has the flavour of a multiple-imputation scheme and does not allow feed-back from the Poisson likelihood for Y_{it} to inform the estimation of N_{iat}.

4.3. Comparisons

In each of the following analyses we consider the model (3)–(5), (8) and first assess how sensitive the ward level residual risks are to the models that we have just described.

Fig. 3(d) compares the use of expected counts based on census populations with an abrupt jump in the year 1985 with those based on populations which are smoothly interpolated via equation (10), i.e. model A. We see that there is very little change in the estimates. We examined the expected counts that result from each approach and found that the differences between these and the observed counts were symmetrically distributed and predominantly lay within ± 0.3.

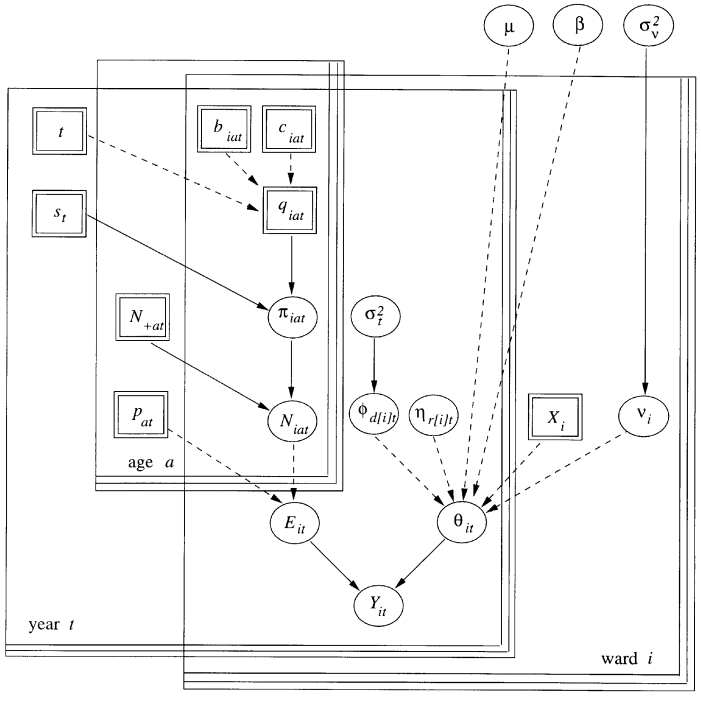

Fig. 5. Directed acyclic graph representation of the final Poisson disease mapping model incorporating registry and district effects and errors in the population denominators

Figs 3(e) and 3(f) compare models A and B and models B and C respectively. Again we see that, in terms of the residual relative risks at least, there are only small differences between the models. As a final check on the sensitivity, we increase the prior variance on π_{iat} by replacing $s_t = 7500$ by $s_t = 3750$. Fig. 3(g) compares the residual relative risk estimates under the two priors and shows that they remain virtually unchanged. We also examined maps of the standard errors of the residual relative risks and found them to be virtually unchanged across the various analyses.

Fig. 6 shows the effect of the various models on the regression parameter measuring the effect of the Carstairs deprivation score. The relative risk that is plotted represents the ratio of risk in an affluent ward (fifth percentile of scores across the region) *versus* the risk in a deprived ward (95th percentile). Ignoring the rightmost point for the moment we see that each of the estimates are greater than 1, indicating that within affluent wards the risk is greater, in line with previous observations (e.g. Henderson *et al.* (1996)). This relationship is largely due to reproductive risk factors; for example early pregnancy is thought to be protective. Fig. 6 shows a clear difference in the estimates based on the pre- and post-1985 data; when all years are combined, progressive refinements to the model lead to increases in both the magnitude and the uncertainty of the deprivation effect. The rightmost point in Fig. 6 shows the relative risk for the odds ratio model (9). This estimate represents the difference in the deprivation effect for breast cancer and for all other cancers. Since for most cancers greater deprivation leads to an increased risk (Jolley *et al.*, 1992) we see a large positive estimate (since β' is large and negative).

Fig. 6 shows the 'attenuation to the null' effect that is well known in covariate

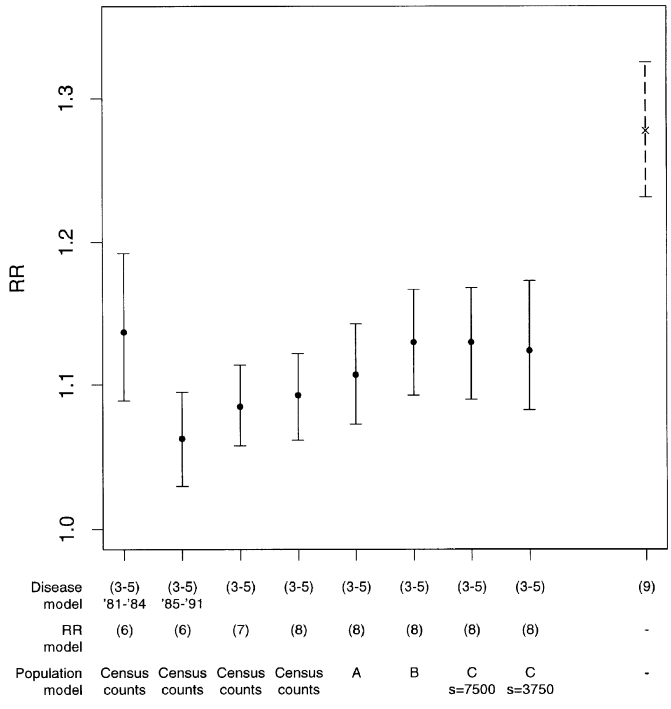

Fig. 6. Regression parameter measuring effect of the Carstairs deprivation score for each model: the quantities plotted are the point estimate and the 95% credible interval for $\exp\{\beta(x_{5\%} - x_{95\%})\}$, i.e. the estimated relative risk of breast cancer in an affluent ward (5th percentile of the Carstairs score) *versus* a deprived ward (95th percentile of the Carstairs score)

measurement error models (e.g. Carroll *et al.* (1995)). However, with the complex model considered here there is no reason why errors in case and population counts will lead to attenuation in general.

We finally carried out an analysis using the most complex model but allowing a separate relative risk parameter $\exp(v_{ij})$, $j = 1, 2$, for each of the periods 1981–1984 and 1985–1991. Fig. 7 shows the ratio of the resulting posterior means of the relative risk estimates in each period by ward. A comparison with Fig. 2 shows that the systematic predominance of ratios less than 1 in the north and greater than 1 in the south has been largely eliminated, suggesting that the lack of agreement in the initial analysis was at least partly due to inaccuracies in the data.

5. Spatial dependence

As stated in Section 3, we would, in general, expect to see spatial dependence in the residual relative risks. To address this issue we attempted to refit the models described in Sections 3 and 4, but allowing for spatially correlated random effects. We utilized model (3)–(5) but with equation (6) replaced by

$$\log(\theta_{it}) = \mu_t + X_i\beta + v_i + u_i, \tag{15}$$

with $v_i \sim_{\mathrm{IID}} N(0, \sigma_v^2)$ and with an intrinsic conditional autoregressive prior for the spatial

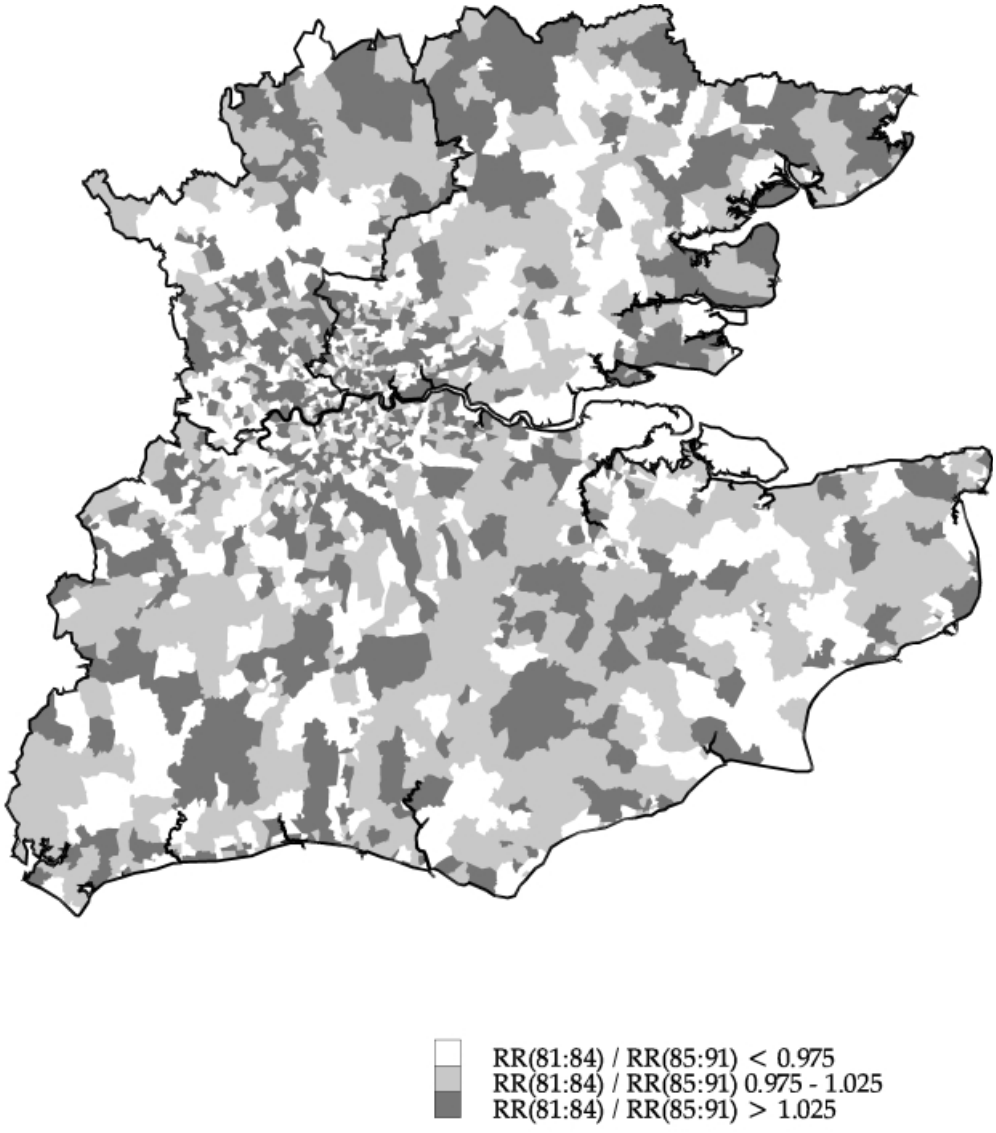

Fig. 7. Ratio of the ward level relative risks in 1981–1984 and 1985–1991 estimated from the final model incorporating registry and district effects and errors in the population denominators: ———, registry boundaries before the amalgamation in 1985

random effects u_i. Specifically we take $u_i|u_j, j \neq i \sim N(\bar{u}_i, \omega_u^2/n_i)$, where $\bar{u}_i$ denotes the mean of the neighbouring wards, where a neighbour is defined to share a common boundary with area i, and n_i denotes the number of neighbours of ward i. Besag *et al.* (1991) advocated the use of this model. The parameter ω_u^2 represents the *conditional* variance of the spatial random effects and is therefore not directly comparable with σ_v^2.

Unfortunately, when we attempted to fit this model including registry effects $\eta_{r[i]t}$, we encountered difficulties since the spatial random effects and the registry effects appeared to be

confounded. To overcome this, we eliminated the need for registry main effects in the linear predictor by obtaining age- and year-specific reference rates separately for each registry and by using these to form the expected numbers. The reference rates are displayed in Fig. 8. As expected, the northern registry rates are lower than the southern rates before the amalgamation in 1985, which we believe represents case underascertainment in the north. This pattern occurs consistently across all age groups apart from the eldest, but the rates for this group are estimated from small numbers and hence have large sampling variability. The effect of the introduction of breast cancer screening in 1989 is evident in the increased incidence in the 50–60-year-old people from this year onwards.

Using the registry-adjusted expected numbers we refitted the basic model (3)–(5) for each time period (1981–1984 or 1985–1991) with either non-spatial random effects (6) or both unstructured and spatial random effects (15). The posterior means of the relative risks from the model with spatial random effects are plotted against their counterparts from the non-spatial model for the corresponding time period in Fig. 9. We see that there are some

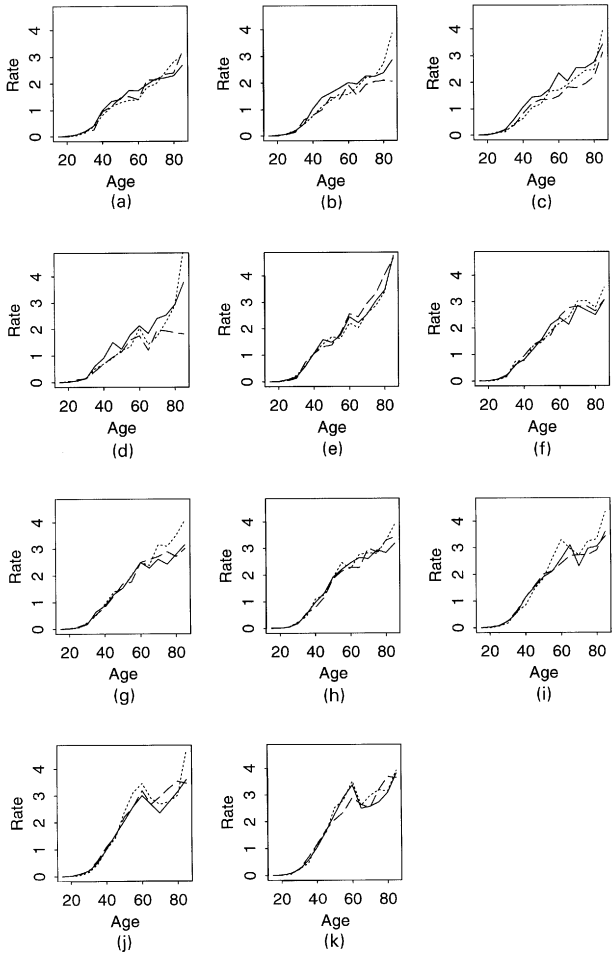

Fig. 8. Registry-specific reference rates (per 1000) by year and age group (— —, North East Thames region; - - - - -, North West Thames region; ———, South Thames region): (a) 1981; (b) 1982; (c) 1983; (d) 1984; (e) 1985; (f) 1986; (g) 1987; (h) 1988; (i) 1989; (j) 1990; (k) 1991

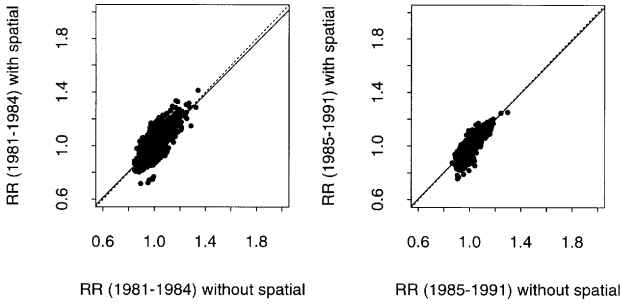

Fig. 9. Posterior means of relative risk parameters obtained from analyses with and without spatial random effects, for the two time periods corresponding to preamalgamation and post-amalgamation

differences, but qualitatively the estimates are in agreement. The posterior means of the regression coefficients β are virtually unchanged when spatial effects are added and the posterior standard deviations were increased slightly. In the non-spatial model the posterior medians of σ_v are 0.15 and 0.10 in the first and second time periods respectively. When the spatial components are added, as expected these standard deviations reduce to 0.10 and 0.07. The empirical standard deviations of the spatial random effects u_i were 0.10 and 0.07, indicating that there is some spatial dependence in these data.

When we added in DHA effects to the spatial model, the posterior medians of σ_t were

$$0.07,\ 0.07,\ 0.05,\ 0.10,$$

for the preamalgamation period, and

$$0.04,\ 0.03,\ 0.02,\ 0.05,\ 0.06,\ 0.03,\ 0.04,$$

for the post-amalgamation period. A comparison with the equivalent parameters in the non-spatial model shows that these parameters are slightly reduced. This could be because we have now adjusted for the registry effects more subtly; we no longer have a single registry-specific adjustment but we allow the reference rates to be estimated separately for each age group. When the DHA effects were added the spatial effects were reduced; the empirical standard deviation of the spatial effects was 0.05 whereas the posterior median of σ_v was 0.08. Again the posterior distribution of β was very similar to that for the equivalent non-spatial analysis.

6. Discussion

In Section 5 we presented an alternative approach to account for registry effects by using a separate set of reference rates for each registry. This would be beneficial in the non-spatial analysis also since it allows for more flexible modelling of registry effects by enabling them to vary across different age groups. We could also follow this approach at the district level, though the sparsity of cases per district would lead to very unstable estimates. In this case the rates could be smoothed across districts by modelling the rates as random effects, and also smoothed across age groups by using a first-order autoregressive prior (e.g. Clayton (1996)).

When modelling the population counts we have assumed that the census and Registrar General's counts are exact, which is clearly not so. It would be straightforward also to assume that these quantities were measured with error, but again, without knowing anything

about the size of the measurement error, we would be essentially carrying out a sensitivity analysis. There are several sources of information that could inform the size of these errors, e.g. local population registries and the 'Estimating with confidence' project (Simpson *et al.*, 1995). In particular the latter found that inaccuracies were greatest in areas that were deprived, contained a relatively young population and contained an ethnically mixed population. As an alternative to modelling the population counts we could take an errors-in-variables approach to modelling the expected counts directly. This approach has the advantage of being more straightforward computationally but it is much more difficult to input prior information concerning the size of the errors in a specific stratum.

Acknowledgements

The authors would like to thank James Bennett, Paul Elliott and Ravi Maheswaran for assistance in preparing the data and for useful discussions, and three referees and the Joint Editor for comments that helped to clarify the ideas of the paper. In particular we appreciate the comments of one referee on the appropriateness of the assumption that is the back-bone of the proportional binomial model described in Section 4.1. The authors would also like to thank the ONS who made the postcoded cancer data available for use. The population data came from the 'Estimating with confidence' project. This work is based on data provided with the support of the Economic and Social Research Council (ESRC) and JISC and uses census and boundary material which are copyright of the Crown, the Post Office and the ED-LINE Consortium. This work was partially funded by the Pan Thames Environmental R&D Programme, project reference 339, and grants from the ESRC (H519255036) and the European Union BIOMED II program (PL96 3488).

References

Bernardinelli, L., Pascutto, C., Best, N. G. and Gilks, W. R. (1997) Disease mapping with errors in covariates. *Statist. Med.*, **16**, 741–752.

Besag, J., York, J. and Mollié, A. (1991) Bayesian image restoration with two applications in spatial statistics. *Ann. Inst. Statist. Math.*, **43**, 1–59.

Bray, I. and Wright, D. E. (1988) Application of Markov chain Monte Carlo methods to modelling birth prevalence of Down syndrome. *Appl. Statist.*, **47**, 589–602.

Breslow, N. and Day, N. E. (1987) *Statistical Methods in Cancer Research*, vol. 2, *The Analysis of Cohort Studies*. Lyon: International Agency for Research on Cancer.

Carroll, R. J., Ruppert, D. and Stefanski, L. A. (1995) *Measurement Error in Nonlinear Models*. London: Chapman and Hall.

Carstairs, V. and Morris, R. (1991) *Deprivation and Health in Scotland*. Aberdeen: Aberdeen University Press.

Clayton, D. G. (1996) Generalised linear mixed models. In *Markov Chain Monte Carlo in Practice* (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter), pp. 279–301. New York: Chapman and Hall.

Clayton, D. G. and Bernardinelli, L. (1992) Bayesian methods for mapping disease risk. In *Geographical and Environmental Epidemiology: Methods for Small-area Studies* (eds P. Elliott, J. Cuzick, D. English and R. Stern), pp. 205–220. Oxford: Oxford University Press.

Clayton, D. G. and Kaldor, J. (1987) Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, **43**, 671–682.

Gulliford, M. C., Bell, J., Bourne, H. M. and Petruckevitch, A. (1993) The reliability of cancer registry records. *Br. J. Cancer*, **67**, 819–821.

Hawkins, M. M. and Swerdlow, A. J. (1992) Completeness of cancer and death follow-up obtained through the National Health Service Central Register for England and Wales. *Br. J. Cancer*, **66**, 408–413.

Henderson, B. E., Pike, M. C., Bernstein, L. and Ross, R. K. (1996) Breast Cancer. In *Cancer Epidemiology and Prevention* (eds D. Schottenfeld and J. F. Fraumeni), 2nd edn, pp. 1022–1039. Oxford: Oxford University Press.

Jolley, D., Jarman, B. and Elliott, P. (1992) Socio-economic confounding. In *Geographical and Environmental Epidemiology: Methods for Small-area Studies* (eds P. Elliott, J. Cuzick, D. English and R. Stern), pp. 115–124. Oxford: Oxford University Press.

Jordan, P., Brubacher, D., Tsugane, S., Tsubono, Y., Gey, K. F. and Moser, U. (1997) Modelling of mortality data from a multi-centre study in Japan by means of Poisson regression with errors in variables. *Int. J. Epidem.*, **26**, 501–507.

Kelsall, J. E. and Wakefield, J. C. (1999) Discussion on Bayesian models for spatially correlated disease and exposure data (by N. G. Best, L. A. Waller, A. Thomas, E. M. Conlon and R. Arnold). In *Bayesian Statistics 6* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), p. 151. Oxford: Oxford University Press.

Mollié, A. (1996) Bayesian mapping of disease. In *Markov Chain Monte Carlo in Practice* (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter). New York: Chapman and Hall.

Office for National Statistics (1997) *1991 Cancer Statistics Registrations*. London: Office for National Statistics.

Office of Population Censuses and Surveys (1991) *Making a Population Estimate in England and Wales*. London: Office of Population Censuses and Surveys.

Parkin, D. M., Chen, V. W., Ferlay, J., Galceran, J., Storm, H. H., Young, J. and Whelan, S. L. (1994) *Comparability and Quality Control in Cancer Registration*. Lyon: International Agency for Research on Cancer.

Pickle, L. W., Mungiole, M., Jones, G. K. and White, A. A. (1996) *Atlas of United States Mortality*. Atlanta: Centers for Disease Control and Prevention.

Richardson, S. (1992) Statistical methods for geographical correlation studies. In *Geographical and Environmental Epidemiology: Methods for Small-area Studies* (eds P. Elliott, J. Cuzick, D. English and R. Stern), pp. 181–204. Oxford: Oxford University Press.

Simpson, S., Tye, R. and Diamond, I. (1995) What was the real population of local areas in mid-1991? *Working Paper 10*. Estimating with Confidence Project, Department of Social Statistics, University of Southampton, Southampton.

Spiegelhalter, D. J. (1998) Bayesian graphical modelling: a case-study in monitoring health outcomes. *Appl. Statist.*, **47**, 115–133.

Spiegelhalter, D. J., Thomas, A. and Best, N. G. (1998) *WinBUGS User Manual, version 1.1.1*. Cambridge: Medical Research Council Biostatistics Unit. (Available from `http://www.mrc-bsu.cam.ac.uk/bugs`.)

Swerdlow, A. J. (1986) Cancer registration in England and Wales: some aspects relevant to interpretation of the data. *J. R. Statist. Soc.* A, **149**, 146–160.

Swerdlow, A. and dos Santos Silva, I. (1993) *Atlas of Cancer Incidence in England and Wales, 1968-85*. Oxford: Oxford University Press.

Wakefield, J. C. and Elliott, P. (1999) Issues in the statistical analysis of small area health data. *Statist. Med.*, to be published.

York, J., Madigan, D., Heuch, I. and Lie, R. T. (1995) Birth defects registered by double sampling: a Bayesian approach incorporating covariates and model uncertainty. *Appl. Statist.*, **44**, 227–242.